# Frequent closed itemsets based condensed representations for association rules

Nicolas Pasquier

# Chapter XIII
# Frequent Closed Itemsets Based Condensed Representations for Association Rules

**Nicolas Pasquier**
~~*Université de Nice, France*~~

*Laboratoire I3S, University of Nice Sophia-Antipolis / CNRS (UMR6070), France*

## ABSTRACT

*After more than one decade of researches on association rule mining, efficient and scalable techniques for the discovery of relevant association rules from large high-dimensional datasets are now available. Most initial studies have focused on the development of theoretical frameworks and efficient algorithms and data structures for association rule mining. However, many applications of association rules to data from different domains have shown that techniques for filtering irrelevant and useless association rules are required to simplify their interpretation by the end-user. Solutions proposed to address this problem can be classified in four main trends: constraint-based mining, interestingness measures, association rule structure analysis, and condensed representations. This chapter focuses on condensed representations that are characterized in the frequent closed itemset framework to expose their advantages and drawbacks.*

## INTRODUCTION

Since the definition of association rules in the early 1990's by Agrawal *et al.* (1993), intensive studies have been conducted to produce efficient association rule mining algorithms from large datasets. Association rules were defined as conditional rules depicting relationships between occurrences of attribute values, called *items*, in data lines[1]. An association rule $A \rightarrow C$ states that a significant proportion of data lines containing items in the *antecedent* $A$ also contain items in the *conse-*

*quent C*. The *support* of a rule is the proportion, or number, of data lines containing all items in the rule to assess the scope, or frequency, of the rule. The confidence of a rule is the proportion of data lines containing the consequent among data lines containing the antecedent. The task of association rule mining consists in discovering all rules with support at least equal to the user defined minimum support threshold *minsup* and that have a confidence at least equal to the user defined minimum confidence threshold *minconf*. Such rules are called *valid* or *strong* association rules.

This approach to association rule mining suffers from several well-known drawbacks described in many researches and application reports (Brijs *et al.*, 2003). The first of these problems is the difficulty to define appropriate *minsup* and *minconf* thresholds. Choosing too high values may lead to the miss of important relations corresponding to association rules with support lower than *minsup*. Choosing too low values may lead to performance problems, as more itemsets are frequent, and to extract association rules that are irrelevant or useless because of their limited scope. The second problem is related to the confidence measure used to assess the precision of the rule. This measure does not consider the frequency, or support, of the consequent of the rule and thus, an association rule can exhibit a relation between two statistically uncorrelated itemsets (Brin *et al.*, 1997). The third problem is related to the huge number of association rules generated in most cases. This number can range from several thousands to several millions and the set of association rules can be difficult to manage and interpret (Toivonen *et al.*, 1995; Bayardo *et al.*, 2000). The fourth problems is related to the presence of many redundant association rules in the result. Redundant association rules are rules which information is contained in other rules and that can thus be deduced from them (Matheus *et al.*, 1993). These rules do not bring additional knowledge to the user and should be removed from the result as they lower the result's accuracy and relevance and harden the management and interpretation of extracted rules.

The frequent closed itemsets framework was introduced to address the efficiency problem of association rule mining from dense and correlated data. Several posterior researches have shown that this framework is also well-fitted to address the problem of redundant association rules filtering. We first present association rule mining and frequent itemsets and frequent closed itemsets frameworks. Then, we briefly review several approaches proposed to address the four problems of association rule mining mentioned above. Then, we describe condensed representations and bases for association rules and characterize them in the frequent closed itemsets framework to show their advantages and drawbacks.

## ASSOCIATION RULE MINING

In order to improve the extraction efficiency, most algorithms for mining association rules operate on binary data represented in a transactional or binary format. This also enables the treatment of mixed data types, resulting from the integration of multiple data sources for example, with the same algorithm. The transactional and binary representations of the example dataset *D*, used as a support in the rest of the chapter, are shown in Table 1. In the transactional or *enumeration* format represented in Table 1(a) each object, called *transaction* or *data line*, contains a list of items. In the binary format represented in Table 1(b) each object[2] is a *bit vector* and each bit indicates if the object contains the corresponding item or not.

An itemset is a lexicographically ordered set of items and the *support* of an itemset $A$ is the proportion, or number, of objects containing it: $support(A) = count(A) / count()$ where $count(A)$ is the number of objects containing $A$ and $count()$ is the total number of objects. For example, the support of the itemset $\{b, c\}$, denoted $bc$ for short,

*Table 1. Dataset Representations.*

| Object | Items |
|---|---|
| 1 | a b c e |
| 2 | b c e |
| 3 | a b c e |
| 4 | a c d |
| 5 | b c e |

(a) Enumerations

| **Object** | ***a*** | ***b*** | ***c*** | ***d*** | ***e*** |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 1 |
| 2 | 0 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 | 0 | 1 |

(b) Bit vectors

is 2/5 in the example dataset *D*. The support of association rule $R: A \rightarrow C$ between two itemsets *A* and *C* is $support(R) = support(A \cup C)$ and it confidence is $confidence(R) = support(A \cup C) / support(C)$. For instance, in dataset *D* we have $support(a \rightarrow bc) = support(abc) = 2/5$ and $confidence(a \rightarrow bc) = support(abc) / support(a) = 2/3$. Association rule mining algorithms can be classified according to several criteria: The theoretical framework they are based on, the search space traversal they perform or the data structures they use. The two following sections present the main theoretical frameworks proposed for association rule mining: The *frequent itemsets* and the *frequent closed itemsets* frameworks.

## Frequent Itemsets Framework

The frequent itemsets framework defined by Agrawal *et al.* (1993) is based on the following decomposition of the problem:

1. Extract frequent itemsets, i.e. itemsets that have a support at least equal to *minsup.*
2. Generate association rules between itemsets that have a confidence at least equal to *minconf.*

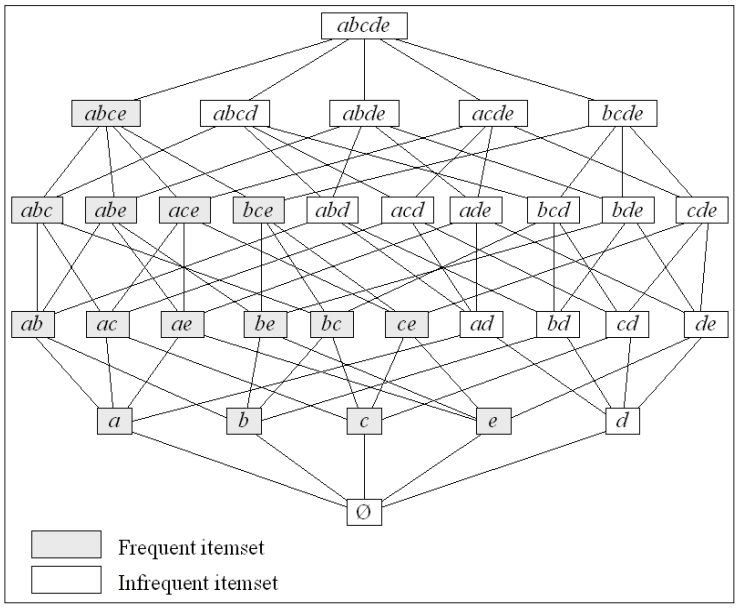
The second phase is straightforward once all frequent itemsets are discovered. However, the search space of the first phase is the itemset lattice, or subset lattice, which size is exponential in the size of set of items and extracting all frequent itemsets was shown to be an NP-Complete problem (Angiulli *et al.*, 2001). The itemset lattice for the example dataset *D* is represented in Figure 1.

This lattice contains $2^{|I|}$ itemsets, where $I = \{a, b, c, d, e\}$ is the set of items and frequent itemsets for *minsup* = 2/5 are outlined. Frequent itemset identification requires lattice traversal and is thus computationally expensive. Optimized traversals of the search space and efficient data structures and implementation techniques are required to obtain acceptable response times. The frequent itemset approach was developed for extracting association rules from very large datasets containing weakly correlated data, such as market basket data. However, this approach faces important problems of efficiency with dense or highly correlated data as the number of frequent itemsets, and by the sequel of association rules, can be very large (Brin *et al.*, 1997). A recent review of association rule mining algorithms in these three trends can be found in Ceglar & Roddick (2006).

## Frequent Closed Itemsets Framework

The frequent closed itemsets framework was introduced in Pasquier *et al.* (1998) with the Close algorithm to address the problem of association rule mining from dense datasets (Pasquier *et al.*,

*Figure 1. Itemset Lattice.*



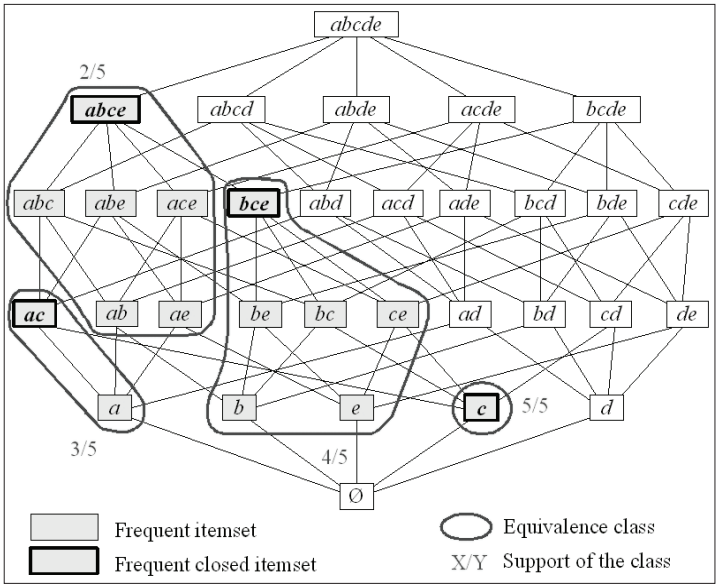

1999a). This framework is based on the closure operator of the Galois connexion used in Formal Concept Analysis (Ganter *et al.*, 2005). It defines frequent closed itemsets that constitute a minimal representation for frequent itemsets. The Galois closure $\gamma(A)$ of an itemset $A$ is the intersection of all objects containing $A$ and a closed itemset is an itemset that is equal to its closure. All frequent itemsets and their supports can be straightforwardly deduced from frequent closed itemsets. Algorithms based on the frequent closed itemset extraction use the following decomposition of the problem:

1. Extract frequent closed itemsets, i.e. closed itemsets that have a support at least equal to *minsup.*
2. Generate association rules that have a confidence at least equal to *minconf* from frequent closed itemsets.

The search space of the first phase is the closed itemset lattice and the potentially frequent closed itemsets for the example dataset $D$ are outlined in Figure 3. A closed itemset is a maximal[3] set of items common to a set of objects. For instance, *ac* is a closed itemset since it is the maximal set of items common to objects 1, 3 and 4, that is their intersection. Frequent closed itemsets for the example dataset $D$ and $minsup = 2/5$ are outlined in Figure 2. These frequent closed itemsets, with their respective supports, summarize all frequent itemsets and their supports (Pasquier *et al.*, 1998). See Ben Yahia *et al.* (2006), Pasquier (2005) and Valtchev *et al.* (2004) for reviews of frequent closed itemset based algorithms.

These algorithms use the closure property to exclude from the search space traversal many itemsets that are useless for association rule construction: The non closed frequent itemsets. Since the closure $\gamma(A)$ of an itemset $A$ is the intersection of all objects containing $A$, the support of $A$ is equal to the support of $\gamma(A)$. All itemsets are contained and have the same support as their closure, that is their first closed superset. For instance, in Figure 2, itemset *b*, *e*, *be*, *bc* and

*Figure 2. Closed Itemsets and Equivalence Classes.*




*ce* are deducible with support from the closed itemset *bce* and by consequence are useless for association rule mining. All itemsets with the same closure form an *equivalence class* (Bastide *et al.*, 2000b). Each equivalence class contains a unique maximal closed itemsets that is the closure of itemsets in the class.

Frequent closed itemset based algorithms eliminate from the search all subsets of identified frequent closed itemsets. However, improvements offered by this approach depend on data density and correlation. Experiments conducted on market basket data showed no efficiency improvement while experiments on census and biological data showed improvements both in execution times and memory usage. This is due to the sparse and weakly correlated nature of market basket data and the dense and correlated nature of census and biological data (Brin *et al.*, 1997; Pfaltz & Taylor, 2002). Despite these improvements, efficient search space traversals, data structures and implementation techniques are required to obtain acceptable response times when large high-dimensional datasets are mined. Recent algorithms based on the frequent closed itemset approach have shown important improvements in response times and memory usage, increasing the capability to mine association rules from high-dimensional datasets with very low *minsup* and *minconf* values.

## ASSOCIATION RULES FILTERING METHODS

Association rule mining has been applied in a wide variety of domains such as marketing, finance, telecommunications, bioinformatics, natural sciences and web usage. All these applications have highlighted important problems that often arise when extracted association rules are analyzed and interpreted: The problems of usability and relevance of extracted rules. This problem results mainly from both the number of rules generated and the presence of numerous redundancies among them. Indeed, in most applications several

thousands, and sometimes millions, of association rules are generated. Moreover, the analyze of these rules shows that most often several rules were generated from the same objects and contain similar information. A small subset of these rules can summarize all information contained in this set of rules as other rules can be considered redundant and their suppression does not reduce information (Padmanabhan & Tuzhilin, 2000; Stumme *et al.*, 2001). This problem is crucial when data are dense or correlated since redundant rules may represent the majority of extracted rules in such data (Bastide *et al.*, 2000a; Zaki, 2000). Identifying and extracting only the most informative rules, from the user's viewpoint, among all rules has then become a crucial problem to improve usefulness of extracted association rules (Bayardo & Agrawal, 1999). Solutions proposed to address this problem can be classified in four main trends: The integration of constraints to select a subset of rules, the use of interestingness measures to evaluate the relevance of rules, the comparison of rules to filter similar ones and the extraction of condensed representations. The following subsections give a short overview of constraint-based mining, interestingness measures and association rule structure analysis approaches. Condensed representations for association rules are studied in more details in next the section.

## Constraint-Based Mining

User specified constraints restrict combinations of items allowed to participate to association rules. They reduce the number of extracted association rules according to structural criteria corresponding to the user's preferences. These preferences are defined either as *templates* or as *item constraints*. Templates are rules containing boolean operators that define which items or combinations of items are allowed or forbidden in the antecedent and the consequent of extracted rules (Klemettinen *et al.*, 1994). In a post-processing phase, rules that do not match the user defined templates are discarded. A template definition language, that is an extension of an SQL operator for extracting association rules from relational databases, was proposed in Baralis & Psaila (1997). A performance evaluation of the template approach is presented in Li *et al.* (2004). Item constraints are boolean expressions over the presence or absence of items in frequent itemsets and then association rules (Srikant *et al.*, 1997). These constraints are integrated during the frequent itemset discovery process to discard itemsets that do not match constraints and thus reduce the search space of frequent itemsets. An algorithm integrating item constraints in a depth-first frequent itemset mining approach was proposed in Lu *et al.* (2005).

Item constraints based approaches can naturally take advantage of structures describing hierarchical relations between items such as taxonomies or *is-a* hierarchies. In such hierarchies, items can be generalized in different ways and to different levels of abstraction according to hierarchical relations. Rules between items at different levels of abstraction, called *multi-level* (Srikant & Agrawal, 1995) or *generalized association rules* (Han & Fu, 1995), can then be generated. This approach can significantly reduce the number of extracted association rules as each higher-level rule may summarize several lower-level rules that can be derived given the corresponding item supports (Srikant & Agrawal, 1995). Moreover, with such hierarchies, item constraints can be defined at different levels of abstraction to simplify their definition when the items involved have common ancestors in the hierarchies.

Other general constraints can be integrated in the mining process. These constraints can be classified in *monotonic* and *anti-monotonic* constraints considering their impact on the mining process. Anti-monotonic constraints, such as the frequency constraint, reduce the search-space and can be pushed deep in the discovery process to optimize it (Bayardo *et al.*, 2000). Monotonic constraints, such as domain of values or class of items, can be checked once and no further

checking is required for subsequent phases (Ng *et al.*,1998; Lakshmanan *et al.*, 1999). The problem of mining frequent itemsets satisfying a conjunction of anti-monotonic and monotonic constraints was studied in Pei *et al.* (2001), Boulicaut & Jeudy (2002) and Bonchi *et al.* (2005). Several studies concerned the integration of general constraints in frequent itemsets based algorithms (Pei & Han, 2002; Cheung & Fu, 2004; Leung *et al.*, 2002) and in frequent closed itemset based algorithms (Bonchi & Lucchese, 2004; Bonchi & Lucchese, 2006).

## Interestingness Measures

The use of interestingness measures to assess statistical significance and operational value of rules was proposed in Piatetsky-Shapiro (1991). This approach was studied in a more general context, to select the most relevant association rules, in Toivonen *et al.* (1995). Interestingness measures are usually classified in *objective* and *subjective* measures. Objective measures assess the interestingness of rules according to a statistical significance criterion. Subjective measures compare rules with user's prior knowledge to assess the interestingness of rules according to *unexpectedness* and *actionability* criteria. Unexpected rules either contradict user's beliefs or represent previously unknown relations. Actionable rules are rules the user can act upon to his advantage. See McGarry (2005) for a review of objective and subjective interestingness measures for knowledge discovery. Recent studies showed that combining objective and subjective measures is required to select the most interesting rules: Objective measures first filter potentially interesting rules and then subjective measures select truly interesting rules (Carvalho *et al.*, 2005).

The support and confidence objective measures were introduced to evaluate the association rules interestingness from a statistical viewpoint (Brijs *et al.*, 2003). The use of statistical measures to assess rules' syntactic similarity and prune similar rules was introduced in Piatetsky-Shapiro & Matheus (1994) and Toivonen *et al.* (1995). The use of other statistical measures to overcome support and confidence weakness, that is particularly important in dense correlated data, was suggested in Brin *et al.* (1997). This study was extended in (Silverstein *et al.*, 2000). Hilderman & Hamilton (1999) introduced and evaluated twelve heuristics for rule mining based on measures from information theory, statistics, ecology, and economics. This work was extended, considering four other measures, by Hilderman & Hamilton (2001). Hilderman & Hamilton (2003) also evaluated the combination of twelve objective measures with taxonomic hierarchies for mining generalized association rules. Tan *et al.* (2002) showed that no objective measure is more appropriate than others in all situations. They evaluated twenty objective measures, highlighting several of their key properties, to help the user choosing measures that are well-fitted to the application domain and his expectations. According to Freitas (1999), several factors, such as disjunct size, imbalance of class distributions, attribute interestingness, misclassification costs and asymmetry, should be considered additionally to the traditional coverage[4], completeness and confidence factors of objective measures. He also proposed a generic criterion taking into account all these factors. *Exception rules* are rules that contradict other rules with high support and confidence containing related items. Exception rules can be identified by using an induction algorithm and contingency tables to determine rules' deviation (Liu *et al.*, 1999) or objective measures based on relative entropy (Hussain *et al.*, 2000). See Geng & Hamilton (2006) and Lenca *et al.* (2008) for recent reviews on objective interestingness measures for association rules.

Many subjective measures to assess the interestingness of association rules by comparison with the user's prior knowledge were proposed. The use of background knowledge, such as user's beliefs, to identify *unexpected* association rules

was introduced in Silberschatz & Tuzhilin (1996). In this approach, user's beliefs are defined in a knowledge base used during the mining process in combination with an unexpectedness heuristic to estimate rules' interestingness from the user's beliefs viewpoint. In Padmanabhan & Tuzhilin (1998), user's beliefs are represented in the same format as association rules and only rules that contradict existing beliefs are mined. In Liu *et al.* (1997), the user defines *general impressions* that express positive or negative relations between items. Association rules are then compared to this knowledge to rank *unexpected* or *confirming* rules. A confirming rule is a rule that matches user's beliefs. In Liu *et al.* (1999), the user's background knowledge is expressed in fuzzy rules and *unexpected*, *confirming* or *actionable* rules are extracted. In Wang *et al.* (2003), the user defines a preference model that characterizes how background knowledge should be used to evaluate rules' unexpectedness. Then, an algorithm extracts unexpected rules satisfying user defined thresholds of minimum *unexpectedness significance* and *unexpectedness strength*. In Jaroszewicz & Scheffer (2005), the interestingness of a rule is evaluated by comparing the difference between its support in the dataset and in a Bayesian network expressing user's prior knowledge. See Geng & Hamilton (2006) for a recent review of subjective measures of interestingness.

## Association Rule Structure Analysis

Approaches in this category analyze the structure of association rules to suppress those containing information represented in other association rules. Each association rule is compared to all other rules and is suppressed if it is "similar" to another rule according to the items in its antecedent and its consequent.

Toivonen *et al.* (1995) proposed an algorithm to prune association rules by keeping only association rules with the minimal antecedent. In this approach, if the antecedent $A$ of an association rule $R$: $A \rightarrow C$ is a superset of the antecedent $A'$ of an association rule $R'$: $A' \rightarrow C'$ with the same consequent, then $R$ is suppressed. A similar approach was proposed in Liu *et al.* (1999) with the difference that the association rule $R$ is suppressed if it does not show a positive correlation according to a $\chi^2$ test with respect to $R'$. Padmanabhan & Tuzhilin (2000) combined an algorithm for mining association rules with minimal antecedent with an unexpectedness measure to mine a minimal set of unexpected association rules. However, the supports, precision measures and objects covered by $R$ and $R'$ are not taken into account by these methods. Thus, the precision measures of suppressed association rules cannot be deduced from the resulting set.

The extraction of A-maximal association rules to reduce the number of extracted association rules was proposed in Bayardo & Agrawal (1999). A-maximal association rules are association rules with maximal antecedent among all rules with the same support and the same consequent. Constraints defining which items are allowed, required or forbidden in the extracted A-maximal association rules can also be integrated to further reduce their number.

Succinct association rules defined in Deogun & Jiang (2005) are strong association rules filtered using a strategy based on a model called MaxPUF for maximal potentially useful association rules. As for other approaches in this category, the resulting set is not lossless since the capability to deduce all strong association rules with their statistical measures is not ensured.

# CONDENSED REPRESENTATIONS

In this section, we characterize condensed representations for association rules using the frequent closed itemsets framework. To simplify this characterization, we distinguish two classes of association rules: *Approximate* or *partial association rules* that have a confidence less than 100

% and *exact association rules* that have a 100 % confidence. Approximate association rules have some counter-examples in the dataset whereas exact association rules have no counter-example in the dataset.

A condensed representation is a reduced set of association rules that summarizes a set of strong association rules. Strong association rules designate association rules extracted using a classical frequent itemset approach, that is all association rules with statistical measures, computed from itemset supports, at least equal to the user defined thresholds. A condensed representation is *lossless* if all strong association rules can be deduced from it. Information lossless condensed representations are called *generating sets*. Generating sets that are minimal with respect to the number of association rules are called *minimal covers* or *bases for association rules* (Pasquier *et al.*, 1999b; Zaki, 2000). A basis for association rules is thus a condensed representation with the two following properties:

1. **Non-redundancy**: A basis contains no redundant rule according to the inference rules[5] considered. That means each association rule in the basis cannot be deduced if suppressed from the basis. In other words, each association rule of the basis must contain information not deducible from other association rules of the basis.
2. **Generating set**: A basis enables the inference of all strong association rules according to the set of inference rules considered. That means all strong association rules can be deduced from the basis.

Condensed representations and bases for association rules were defined to bring to the end-user a set of association rules as small as possible. Bases are condensed representations with more restrictive properties: They are minimal sets of association rules from which all strong association rules can be deduced by inference. This deduction, or inference, relies on a set of inference rules defining which association rules can be deduced from other association rules and are thus redundant. A set of inference rules is called an *inference system*.

## Inference Systems

In the domains of databases and data analysis, such as Formal Concept Analysis, the closure of a set $S$ of implication rules according to an inference system is the set $S+$ of all implication rules that can be inferred from $S$. Then, an implication rule is redundant if its suppression does not change this closure. Armstrong's axioms are inference rules proposed in the field of database conception for generating the closure of a set of functional dependencies between attribute values (Armstrong, 1974). Armstrong's inference system, recalled in Figure 3, is well-suited for association rules as implication rules are closely related to functional dependencies in the database domain (Maier, 1983; Valtchev *et al.*, 2004). Indeed, exact association rules are implication rules that are

*Figure 3. Armstrong's Axioms.*

| | |
|---|---|
| Reflexivity | $: X \supseteq Y \vdash X \Rightarrow Y$ |
| Augmentation | $: X \Rightarrow Y \vdash XZ \Rightarrow YZ$ |
| Transitivity | $: X \Rightarrow Y \wedge Y \Rightarrow Z \vdash X \Rightarrow Z$ |
| Union | $: X \Rightarrow Y \wedge X \Rightarrow Z \vdash X \Rightarrow YZ$ |
| Decomposition | $: X \Rightarrow YZ \vdash X \Rightarrow Y \wedge X \Rightarrow Z$ |
| Pseudo-transitivity | $: X \Rightarrow Y \wedge WY \Rightarrow Z \vdash XW \Rightarrow Y$ |

frequent enough in the dataset, according to the user defined frequency threshold *minsup*, and approximate association rules are partial implication rules that are frequent enough in the dataset to be considered useful (Valtchev *et al.*, 2004).

According to these definitions, several bases can be defined depending on the inference system considered. This inference system determines which association rules are considered redundant and are thus suppressed to constitute the basis.

## Redundant Association Rules

To state the problem of redundant association rules, consider the three following rules that can be extracted from the example dataset $D$: $c \rightarrow b$, $c \rightarrow e$, $c \rightarrow be$. These approximate association rules have identical support (4/5) and confidence (4/5) as they are computed from the same objects (1, 2, 3 and 5). Obviously, the information in the two first rules is summarized by the third rule and thus, the two first are informatively useless. These two rules can be deduced using the union inference rule of Armstrong's axioms for instance.

Consider now objects 2 and 4 in Table 1 from which the following nine association rules can be extracted: $b \rightarrow c$, $b \rightarrow e$, $b \rightarrow ce$, $bc \rightarrow e$, $be \rightarrow c$, $e \rightarrow c$, $e \rightarrow b$, $e \rightarrow bc$, $ce \rightarrow b$. These exact association rules have a support of 4/5 and a confidence of 100 %. Using the deduction, augmentation and pseudo-transitivity inference rules of Armstrong's axioms, these nine rules can be inferred from the following two rules: $b \rightarrow ce$ and $e \rightarrow bc$. All information contained in the nine rules are contained in these two rules. Moreover, the presence of $c$ in the antecedent of rules $bc \rightarrow e$ and $ce \rightarrow b$ is not significant since their statistical measures do not change if $c$ is removed as $support(\{b\}) = support(\{bc\})$ and $support(\{e\}) = support(\{ce\})$. This reasoning also applies to the presence of $e$ in the antecedent of rule $be \rightarrow c$.

In the database and Formal Concept Analysis domains, a functional dependency or an implication rule is redundant if its suppression does not modify the result of the closure of the set of dependencies or rules according to an inference system. This definition was adapted to association rules and a consensus among researchers is now established to consider that an association rule is redundant if its suppression does not modify the result of the closure of the set of association rules according to an inference system. In other words, a redundant association rule can be inferred from other strong association rules. Extracting only non-redundant association rules reduces as much as possible the number of rules without loosing the capability to retrieve other rules given the inference system. This greatly simplifies their post-processing, that is their management and exploration. In the literature, non-redundant association rules are sometimes called *non-derivable* or *non-deducible* association rules.

In the following, condensed representations for association rules are characterized using the frequent closed itemsets framework. These condensed representations are bases or minimal covers defined according to different sets of inference rules, corresponding to different goals, and thus have different properties with regard to their intelligibility for the end-user.

## Duquenne-Guigues and Luxenburger Bases

The DG Basis (Duquenne & Guigues, 1986) and the Proper Basis (Luxenburger, 1991) for global and partial implications respectively were adapted to the association rule framework in Pasquier *et al.* (1999b). The mathematical, structural and informative properties of these bases was studied in several research papers (Cristofor & Simovici, 2002; Hamrouni *et al.*, 2006; Kryszkiewicz, 2002)

The DG Basis was defined in the context of implication rules and thus does not consider confidence in the inference system used to define redundant rules and deduce all strong rules from the bases. The DG Basis for exact association rules

*Figure 4. DG Inference Rules.*

$$A \Rightarrow B \wedge C \Rightarrow D \vdash AC \Rightarrow BD$$
$$A \Rightarrow B \wedge B \Rightarrow C \vdash A \Rightarrow C$$

is defined by frequent pseudo-closed itemsets. A frequent pseudo-closed itemset *A* is a non-closed itemset that includes the closures of all frequent pseudo-closed itemsets included in *A*. This basis contains all association rules between a frequent pseudo-closed itemset and its closure. The DG Basis for the example dataset *D* and *minsup* = 2/5 and *minconf* = 2/5 is represented in Table 2.

The DG Basis for exact association rules is a minimal set, with respect to the number of extracted exact rules. All strong exact association rules can be deduced from the DG Basis using the inference rules given in Figure 4.

The DG Basis is the minimal generating set with respect to the number of rules for a set of implication rules (Ganter *et al.*, 2005). The same property was demonstrated for the DG Basis for association rules (Pasquier *et al.*, 1999b). However, statistical measures of all strong association rules inferred cannot be deduced from the DG Basis. Association rules inferred using the first inference rule can have inferior support compared to the rules used for inferring them. Frequent closed itemset supports are then necessary to deduce statistical measures of all exact association rules from the DG Basis (Cristofor & Simovici, 2002; Kryszkiewicz, 2002).

The Proper Basis for association rules contains all association rules between two frequent closed itemsets related by inclusion. This basis contains exactly one association rule for each pair of equivalence classes which frequent closed itemsets are related by inclusion. The Proper Basis for the example dataset *D* and *minsup* = 2/5 and *minconf* = 2/5 is represented in Table 3.

The transitive reduction of the Proper Basis (Pasquier *et al.*, 1999b) is the reduction of the Proper Basis according to the transitivity inference rule such as defined in Armstrong's axioms. This inference rule states that given three frequent closed itemsets *A*, *B* and *C* such that $A \supset B \supset C$, the confidence of the association rule $A \rightarrow C$ can be computed from the confidences of association rules $A \rightarrow B$ and $B \rightarrow$ C: *confidence*($A \rightarrow C$) = *confidence*(A $\rightarrow B$) × *confidence*($B \rightarrow C$). Then, $A \rightarrow$ C is called a transitive associa-

*Table 2. DG Basis for Exact Association Rules.*

| Pseudo-closed itemset | Closure | Association rule | **Support** |
|---|---|---|---|
| {*a*} | {*ac*} | $a \rightarrow c$ | 3/5 |
| {*b*} | {*bce*} | $b \rightarrow ce$ | 4/5 |
| {*e*} | {*bce*} | $e \rightarrow bc$ | 4/5 |

*Table 3. Proper Basis for Approximate Association Rules.*

| Subset | Superset | Association rule | **Support** | **Confidence** |
|---|---|---|---|---|
| [*ac*] | [*abce*] | $ac \rightarrow be$ | 2/5 | 2/3 |
| [*bce*] | [*abce*] | $bce \rightarrow a$ | 2/5 | 2/4 |
| [*c*] | [*ac*] | $c \rightarrow a$ | 3/5 | 3/5 |
| [*c*] | [*bce*] | $c \rightarrow be$ | 4/5 | 4/5 |
| [*c*] | [*abce*] | $c \rightarrow abe$ | 2/5 | 2/5 |

tion rule. Considering rules in Table 3, the only transitive rule is c ⟶ *abe* and *confidence*(*c* ⟶ *abe*) = confidence(c ⟶ *be*) × *confidence*(*bce* ⟶ a) = 4/5 × 2/4 = 2/5. Another transitivity allows the inference of its confidence: *confidence*(*c* ⟶ *abe*) = confidence(*c* ⟶ *a*) × *confidence*(*ca* ⟶ *be*) = 3/5 × 2/3 = 2/5. This inference can be extended to all statistical measures of precision[6] computed using only supports of the itemsets included in the association rule.

These union of the DG Basis for exact association rules and the transitive reduction of the Proper Basis for approximate association rules is a minimal basis for association rules: No smaller set allows the deduction of the antecedent and consequent of all strong exact and approximate association rules (Pasquier *et al.*, 1999b).

## Informative Bases

Bases defined in the database and data analysis domains, such as the DG Basis and the Proper Basis for implication rules, are minimal in their number of rules given an inference system. However, these bases do not consider the problem of exploration and interpretation of extracted rules by the end-user. They were defined for an automated treatment such as the computation of a database conceptual schema with normalization properties or to have a small set of rules that is easier to manage and treat, and from which all strong rules can be inferred on demand. They were not defined for human reasoning on the information they contain when the end-user explores them. However, in most data mining applications, the interpretation of extracted patterns by the end-user is a crucial step to maximize result's profitability. Consequently, the capability to get all information contained in the set of strong association rules should be considered as a criterion when defining which rules are presented to the end-user. To ensure this capability, the deduction of information contained in suppressed association rules must be natural (Goethals *et al.*, 2005). Thus, to improve the relevance of the set of extracted association rules from the end-user interpretation viewpoint, a basis must possess the two additional properties:

1. **Itemset covering**: Association rules in the basis must cover all combinations of items covered by the set of all strong rules. That means all items contained in an association rule must be contained in an association rule of the basis, possibly containing other items.
2. **Objects covering**: Association rules in the basis must cover all sets of objects covered by the set of all strong rules. That means all association rule representing relationships between items contained in two given sets of objects must be deducible from an association rule of the basis concerning the same two sets of objects.

Bases with these two properties are called *informative bases* and their association rules are called *informative association rules* (Gasmi *et al.*, 2005; Hamrouni *et al.*, 2008; Pasquier *et al.*, 2005). Informative bases are condensed representations that are minimal generating sets of association rules bringing to the user all information about itemset co-occurrences as the set of all strong association rules. By information, we refer to both relationships between sets of items and statistical measure values that assess the frequency and the strength or precision of this relationship. The minimality of informative bases is ensured by eliminating redundant association rules containing information contained in other association rules. They are useless for the end-user interpretation and can be suppressed without reducing information. An inference system derived from Armstrong's axioms to define the informative bases was proposed by Cristofor & Simovici (2002). To characterize the properties of informative bases, the equivalence class framework is presented in the following section.

Kryszkiewicz (2002) demonstrated that the Proper Basis and the DG Basis, if supports of association rules are not considered, are lossless representations of strong exact and approximate association rules respectively. It was also demonstrated using Armstrong's axioms that both the DG Basis and the Proper Basis are not informative.

## Equivalence Classes

Equivalence classes of itemsets are defined using the frequent closed itemsets framework (Bastide *et al.*, 2000b). An equivalence class is a set of itemsets that are contained in the same objects of the dataset, that is they cover the same objects. These itemsets have the same support that is the support of the equivalence class. A frequent equivalence class is an equivalence class which support is greater or equal to the *minsup* threshold. In the following, we refer to an equivalence class by its maximal itemset that is a closed itemset and is unique. For instance, [*bce*] refers to the equivalence class {*b*, *e*, *bc*, *be*, *ce*, *bce*}. Equivalence classes in the example dataset *D* for *minsup* = 2/5 are represented in the lattice of frequent itemsets in Figure 5.

The confidence of an association rules $R: A \rightarrow C$ is $confidence(R) = support(AC) / support(A)$. We deduce that all association rules between two itemsets of the same equivalence class have the same support and a confidence of 100 %. We also deduce that all association rules between two equivalence classes, i.e. between itemsets of the same two equivalence classes, have the same support and the same confidence that is smaller than 100 %. This reasoning can be extended to all precision measures computed from supports of the antecedent and the consequent of the association rule. We show this for respectively exact and approximate association rules in the two following paragraphs.

Consider association rules between two itemsets of the equivalence class [*bce*] such as rule $b \rightarrow ce$, equivalent to $b \rightarrow bce$, between itemsets *b* and *bce*. Since these itemsets belong to the same class they have identical supports and

*Figure 5. Frequent Equivalence Classes.*

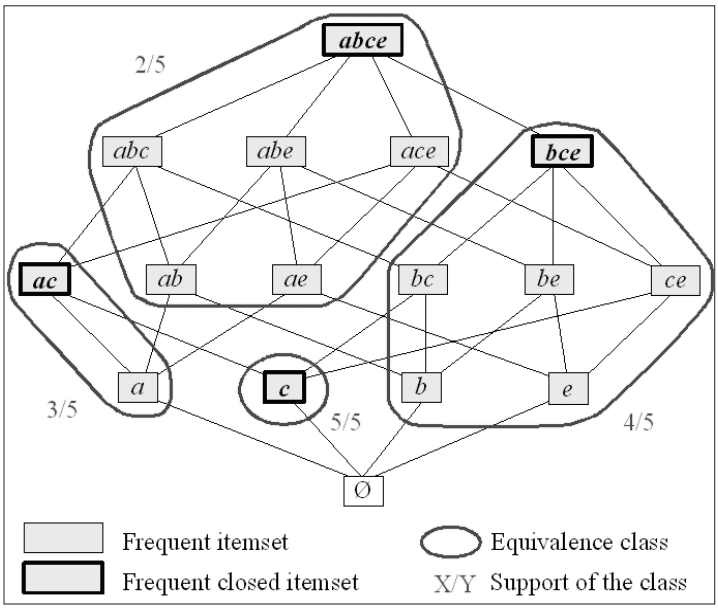
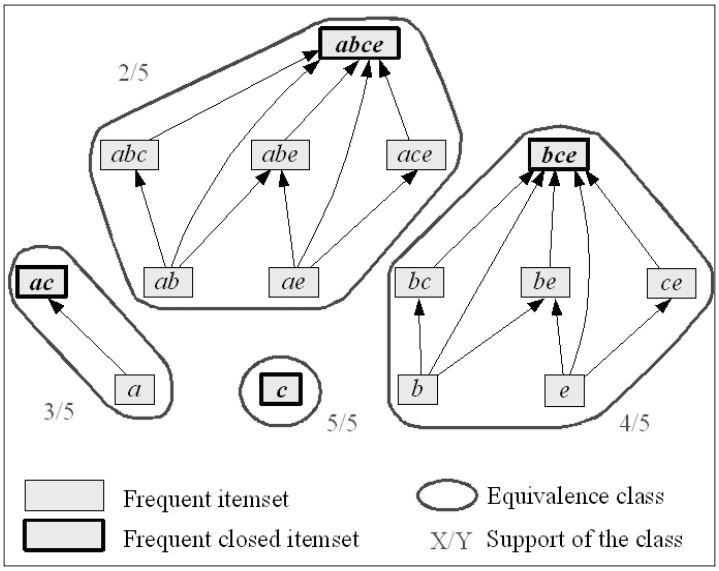


*Figure 6. Exact Association Rules.*




thus, all association rules between two of them have a confidence of 100 % (4/4). The support of these association rules is 4/5 that is the support of the equivalence class. These association rules constitute a *class of exact association rules*. Considering all equivalence classes, i.e. [*c*], [*ac*], [*bce*] and [*abce*], we obtain all strong exact association rules (Pasquier *et al.*, 2005). The nineteen strong exact association rules in the example dataset *D* for *minsup* = 2/5 are represented as directed links in Figure 6. They form three classes of exact association rules since no exact association rule is generated from the equivalence class [*c*] as it contains only the itemset *c*.

Consider now association rules between two itemsets of the equivalence classes [*ac*] and [*abce*] such as rule $a \rightarrow bce$ between *a* and *abce* and rule $ac \rightarrow be$ between *ac* and *abce*. These association rules are represented in Figure 7. Antecedents of these rules are itemsets in the class [*ac*] with *support* = 3/5 and consequents are determined by itemsets in the class [*abce*] with *support* = 2/5. These association rules thus all have a confidence of 2/3 and a support of 2/5. They constitute a *class of approximate association rules*. Each pair of equivalence classes which frequent closed itemsets are related by inclusion, i.e. {[*ac*], [*abce*]}, {[*c*], [*ac*]}, {[*c*], [*bce*]}, {[*c*], [*abce*]} and {[*bce*], [*abce*]}, defines a class of approximate association rules (Pasquier *et al.*, 2005). These five classes of approximate association rules contain the thirty one strong approximate association rules in the

*Figure 7. Approximate Association Rules.*

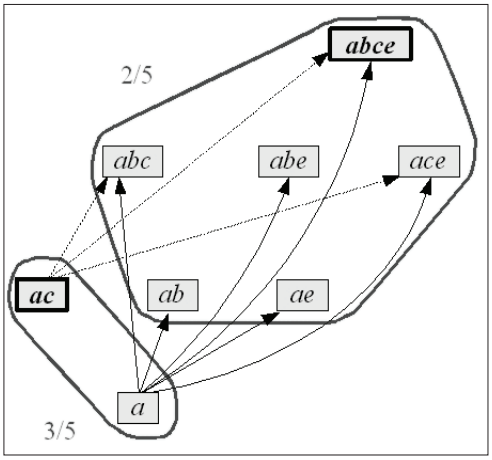
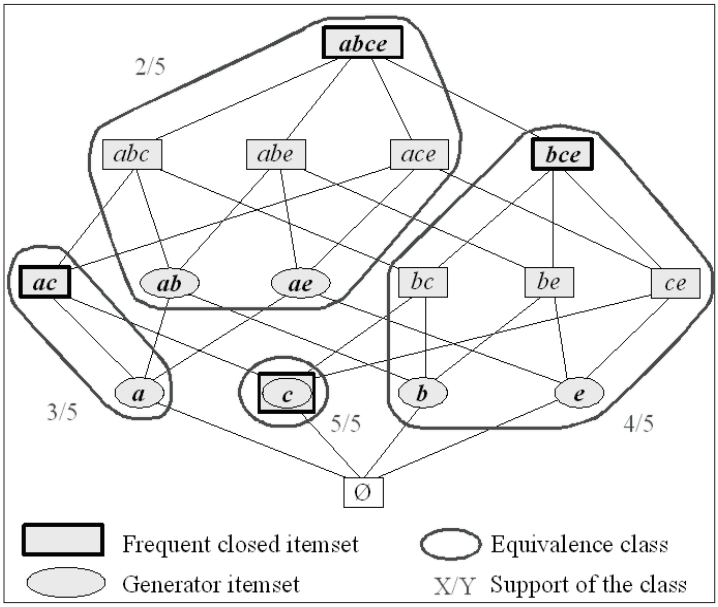


*Figure 8. Frequent Generators.*




example dataset *D* for *minsup* = 2/5 and *minconf* = 2/5.

Each class of exact and approximate association rules regroups all association rules covering the same objects in the dataset. They are thus well-suited to characterize redundant exact and approximate association rules according to the definition of informative bases. Indeed, the space of strong association rules can be divided in classes of association rule and redundant association rules can be characterized inside their class.

Each equivalence class defines an interval of itemsets delimited by the minimal itemsets and the closed itemset of the equivalence class, according to the inclusion relation. These minimal itemsets, called *generators*, have several important properties for the definition of association rule bases.

Generators

Generators of a closed itemset are minimal itemsets which closure is the closed itemset (Pasquier *et al.*, 1998). For instance, the generators of the closed itemset *bce* in the example dataset *D* are *b* and *e*. Itemsets *bc*, *be* and *ce* are not generators of *bce* since they are not minimal in the equivalence class {*b*, *e*, *bc*, *be*, *ce*, *bce*} of *bce* and *c* is a closed itemset and its own unique generator at the same time, since the closure of itemset *c* is *c*. Frequent generators are generators of frequent closed itemsets. Frequent generators in the example dataset *D* for *minsup* = 2/5 and *minconf* = 2/5 are shown in Figure 8.

Frequent generators constitute a relevant generating set for all frequent itemsets with supports and for all strong association rules or condensed representations (Hamrouni *et al.*, 2008; Liu *et al.*, 2007). Moreover, association rules can be computed more efficiently from frequent generators as they are shorter than frequent closed itemsets (Li *et al.*, 2006).

## Minimal Antecedent and Consequent Association Rules

The extraction of the informative basis containing association rules with minimal antecedent and minimal consequent was proposed by Zaki

(2000). In this basis, redundant association rules are filtered inside each class of association rules: Rules with minimal antecedent and minimal consequent among all rules of the class are selected. This filtering is based on the observation that rules with fewer elements in the antecedent are easier to interpret and comprehend (Kryszkiewicz, 1998; Liu *et al.*, 1999; Mc Garry, 2005; Toivonen *et al.*, 1995).

Since generators are the minimal itemsets of an equivalence class, association rules with minimal antecedent in a class are association rules with a generator as antecedent. Considering all classes of association rules, we have a set of association rules covering all sets of items and objects covered by the set of all strong association rules. This basis is informative and lossless: All strong association rules can be deduced with support and precision measures from this set.

Considering classes of exact association rules, association rules with minimal antecedent and consequent are rules between the generators of the class and each of their first supersets in the equivalence class. Exact association rules with minimal antecedent and consequent for the example dataset $D$ and *minsup* = 2/5 and *minconf* = 2/5 are presented in Table 4.

Approximate association rules with minimal antecedent and consequent are rules between a generator $G$ and each of their smallest supersets in another equivalence class which closure $C'$ is a superset of the closure $C$ of $G$. Strong approximate association rules with minimal antecedent and consequent in the example dataset $D$ for *minsup* = 2/5 and *minconf* = 2/5 are presented in Table 5.

*Table 4. Min-Min Exact Association Rules.*

| Generator | Closure | Association rule | **Support** |
|---|---|---|---|
| {*a*} | {*ac*} | *a* → *c* | 3/5 |
| {*ab*} | {*abce*} | *ab* → *c* | 2/5 |
| {*ab*} | {*abce*} | *ab* → *e* | 2/5 |
| {*ae*} | {*abce*} | *ae* → *c* | 2/5 |
| {*ae*} | {*abce*} | *ae* → *b* | 2/5 |
| {*b*} | {*bce*} | *b* → *c* | 4/5 |
| {*b*} | {*bce*} | *b* → *e* | 4/5 |
| {*e*} | {*bce*} | *e* → *b* | 4/5 |
| {*e*} | {*bce*} | *e* → *c* | 4/5 |

Zaki (2004) showed that the number of strong association rules with minimal antecedent and consequent is linear in the number of frequent closed itemsets and that the reduction in the number of association rules will thus be important in most cases. In the example dataset $D$, twenty one minimal antecedent and consequent association rules are strong whereas the set of all strong association rules contains fifty association rules.

Goethals *et al.* (2005) extended this approach to define another condensed representation containing only minimal antecedent and consequent association rules. However, the precision measures of some strong association rules cannot be deduced. It can only be approximated, given a user-specified error bound on the approximation used during the construction of the condensed representation. Thus, contrarily to the basis define by Zaki (2000), this condensed representation is not lossless.

## Generic Bases for Association Rules

*Generic bases for association rules* were defined in Bastide *et al.* (2000a). They were conceived keeping two observations in mind: A basis for association rules presented to the end-user should be informative and rules with fewer elements in the antecedent are easier to interpret and comprehend (Kryszkiewicz, 1998; Liu *et al.*, 1999; Mc Garry, 2005; Toivonen *et al.*, 1995). The Generic bases for exact and approximate association rules contain association rules with minimal antecedent and maximal consequent. In these rules, the difference between the antecedent, that is the premise, and the consequent, that is the conclusion, is maximal and so is their scope as they cover more information. For instance, considering

*Table 5. Min-Min Approximate Association Rules.*

| Generator | Equivalence class | Association rule | **Support** | **Confidence** |
|---|---|---|---|---|
| {*a*} | [*abce*] | $a \rightarrow bc$ | 2/5 | 2/3 |
| {*a*} | [*abce*] | $a \rightarrow be$ | 2/5 | 2/3 |
| {*a*} | [*abce*] | $a \rightarrow ce$ | 2/5 | 2/3 |
| {*b*} | [*abce*] | $b \rightarrow ac$ | 2/5 | 2/4 |
| {*b*} | [*abce*] | $b \rightarrow ae$ | 2/5 | 2/4 |
| {*c*} | [*ac*] | $c \rightarrow a$ | 3/5 | 3/5 |
| {*c*} | [*bce*] | $c \rightarrow b$ | 4/5 | 4/5 |
| {*c*} | [*bce*] | $c \rightarrow e$ | 4/5 | 4/5 |
| {*c*} | [*abce*] | $c \rightarrow ab$ | 2/5 | 2/5 |
| {*c*} | [*abce*] | $c \rightarrow ae$ | 2/5 | 2/5 |
| {*e*} | [*abce*] | $e \rightarrow ab$ | 2/5 | 2/4 |
| {*e*} | [*abce*] | $e \rightarrow ac$ | 2/5 | 2/4 |

approximate association rules $c \rightarrow b$, $c \rightarrow e$ and $c \rightarrow be$ of the same class of association rules, $c \rightarrow be$ is the one with the highest scope as it contains the information contained in each other rule. Statistical measures of differential information use a similar criterion to evaluate rule interestingness (Mc Garry, 2005).

Generators and frequent closed itemsets are respectively the minimal and the maximal itemsets in a class of association rules. Consequently, association rules with minimal antecedent and maximal consequent are association rules between generators and frequent closed itemsets that are their supersets. If all classes of association rules are considered, we obtain a set of association rules covering all sets of items and objects covered by the set of all strong association rules. These rules are the informative association rules with minimal antecedent and maximal consequent. They are called *generic* or *min-max association rules*. They constitute the Generic Basis for association rules (Bastide *et al.*, 2000a).

From a structural viewpoint, an association rule $R$: $A \rightarrow C$ is a generic association rule if there is no association rule $R'$: $A' \rightarrow C'$ with the same support and confidence, whose antecedent $A'$ is a subset of the antecedent $A$ of $R$ and whose consequent $C'$ is a superset of the consequent $C$ of $R$. An inference system based on this definition of generic association rules was proposed in Cristofor & Simovici (2002). Kryszkiewicz (2002) demonstrated that the set of generic association rules is lossless and sound, and that both properties are important since condensed representations of strong association rules that are not sound are of no value even if lossless.

The Generic Basis for exact association rules contains all association rules between a frequent generator $G$ and its closure $C$. Each of these association rules represents several rules within the same class of association rules and covering exactly the same objects and items. Then, the number of generic exact association rules is equal to the number of frequent generators in equivalence classes containing more than one itemset. The Generic Basis for exact association rules for the example dataset $D$ and $minsup = 2/5$ and $minconf = 2/5$ is represented in Table 6.

The Generic Basis for approximate association rules contains all association rules between a frequent generator $G$ and each of the frequent closed itemsets $C_1, ..., C_n$ that are supersets of the

*Table 6. Generic Basis for Exact Association Rules.*

| Generator | Closure | Association rule | **Support** |
|---|---|---|---|
| {a} | {ac} | a → c | 3/5 |
| {ab} | {abce} | ab → ce | 2/5 |
| {ae} | {abce} | ae → bc | 2/5 |
| {b} | {bce} | b → ce | 4/5 |
| {e} | {bce} | e → bc | 4/5 |

closure of *G*. As for exact association rules, each generic approximate association rule represents several approximate association rules in the same class of association rules and covering exactly the same objects and items. The Generic Basis for approximate association rules for the example dataset *D* and *minsup* = 2/5 and *minconf* = 2/5 is represented in Table 7.

The transitive reduction of the Generic Basis for approximate association rules is the reduction according to the transitivity inference rule used to define the transitive reduction of the Proper Basis. Considering association rules in Table 7, the only transitive rule is $c \rightarrow abe$ and we have $confidence(c \rightarrow abe) = confidence(c \rightarrow be) \times confidence(b \rightarrow ace) = 4/5 \times 2/4 = 2/5$. This rule can thus be deduced from rules $c \rightarrow be$ and $b \rightarrow ace$ or $e \rightarrow abc$, or from $c \rightarrow a$ and $a \rightarrow bce$. This rule has a lower confidence than the non-transitive rules as its confidence is the product of their confidences that are lower than 100 %. This transitive reduction allows to further reduce the number of rules by suppressing the generic association rules that have the smallest confidences and are thus the least relevant for the end-user among generic association rules of the same class of association rules.

The union of the Generic bases for exact and approximate association rules is a basis for all strong association rules (Bastide *et al.*, 2000a). The Generic Basis for association rules was defined to provide the end-user with a set of association rules as small as possible, containing only the most relevant association rules and covering all strong co-occurrence relationships between two itemsets. This informative basis holds several properties:

- It contains informative association rules with minimal antecedent and maximal consequent. This property simplifies the interpretation by the end-user, as association rules with smaller antecedents are easier to interpret, and the information in each rule is maximized to minimize their number.
- It implicitly separates approximate association rules and exact association rules. This distinction can be made during the computation to extract a Generic Basis for exact, approximate or both association rules.

*Table 7. Generic Basis for Approximate Association Rules.*

| Generator | Closed superset | Association rule | **Support** | **Confidence** |
|---|---|---|---|---|
| {*a*} | {*abce*} | $a \rightarrow bce$ | 2/5 | 2/3 |
| {*b*} | {*abce*} | $b \rightarrow ace$ | 2/5 | 2/4 |
| {*c*} | {*ac*} | $c \rightarrow a$ | 3/5 | 3/5 |
| {*c*} | {*bce*} | $c \rightarrow be$ | 4/5 | 4/5 |
| {*c*} | {*abce*} | $c \rightarrow abe$ | 2/5 | 2/5 |
| {*e*} | {*abce*} | $e \rightarrow abc$ | 2/5 | 2/4 |

- It is information lossless as all strong association rules can be deduced with their support and precision measures.
- Each generic association rule summarizes a set of association rules covering the same items and objects of the dataset. This property ensures that all information on itemset relationships are in the basis and can be presented to the end-user without having to perform a computation during the interpretation phase.
- Its generation does not require extra computation when a frequent closed itemset based algorithm is used as it can be integrated directly in the mining process.
- It can efficiently be generated from a set of strong association rules as a post-processing phase at little extra computation cost. This generation is straightforward and does not require accessing the dataset or the frequent itemsets. Algorithms for this generation are presented in Pasquier *et al.* (2005).

In the example dataset *D,* fifty association rules are strong for *minsup* = 2/5 and *minconf* = 2/5 whereas the Generic Basis contains eleven association rules. The number of association rules in this basis is linear in the number of frequent closed itemsets. It is not minimal with respect to the number of association rules, but it is a balanced solution for both reducing the number of rules and keeping an easy to interpret set of rules covering all existing relationships between itemsets. The Generic Basis constitutes an interesting starting point to reduce the size of the association rule set without information loss (Hamrouni *et al.*, 2006). Kryszkiewicz (2002) stated that the couple of the Generic bases for exact and approximate association rules combines the ideal properties of a condensed representation for association rules as it is lossless, sound and informative.

## Extensions of the Generic Basis for Association Rules

The problem of defining a generating set that is minimal with respect to the number of association rules was the subject of several recent studies (Dong *et al.*, 2005; Gasmi *et al.*, 2005; Hamrouni *et al.*, 2006). These minimal generating sets are easier to manage and all strong association rules can be deduced from them, with support and precision measures, for interpretation by the end-user. From these generating sets, condensed representations such as the Generic Basis can be generated and presented to the end-user.

Dong *et al.* (2005) showed that some frequent generators can be deduced from other frequent generators by a subset substitution process. A condensed representation named SSMG for *Succinct System of Minimal Generators* from which all generators and all frequent closed itemsets can be inferred with their supports was proposed. In this condensed representation, only one generator, that is the first in the lexicographic order on itemsets, represents each class of association rules. However, Hamrouni *et al.* (2007) demonstrated that the SSMG set is not lossless and new definitions to ensure this lossless property were proposed in Hamrouni *et al.* (2008). A new lossless and sound basis for association rules relying on these definitions was also defined.

Gasmi *et al.* (2005) introduced the *IGB Basis* derived from the Generic bases for exact and approximate association rules. This basis introduces a novel characterization of generic association rules by discerning *factual* and *implicative* association rules instead of exact and approximate ones. Factual association rules have an empty antecedent and highlight unconditional correlations between items whereas implicative association rules have a non empty antecedent. They showed that the IGB basis is lossless and sound and an inference system with conditional

reflexivity, augmentation and decomposition inference rules was defined to enable the inference all strong association rules.

The *Reliable Exact Basis* for exact association rules was defined by Xu & Li (2007) by relaxing the requirements for non-redundancy of the Generic bases and thus suppressing more association rules. This basis is constructed using the certainty factor measure or CF to evaluate generic association rules. The CF of an association rule $R$: $A \rightarrow C$ evaluates both the degree of the belief that the consequent $C$ would be increased if the antecedent $A$ was observed, and the degree of the disbelief that the consequent $C$ would be increased by observing the same antecedent $A$. The Reliable Exact Basis is constructed using the CF to filter association rules in the Generic Basis with measures lower than some user defined thresholds. However, precision measures of all suppressed strong association rules cannot be deduced and this condensed representation is not lossless.

Cheng *et al.* (2008) defined δ-Tolerance Association Rules or δ-TARs using a new concept, called δ-tolerance, based on the approximation of frequent closed itemset supports. An inference system to infer all strong association rules from the δ-TARs, with approximated support and confidence, was also defined. The set of δ-TARs is a condensed representation for the set of strong association rules but is not lossless. The authors affirm that the set of association rules derived from the δ-TARs by this inference system is sound and complete and that approximations of supports and confidences are accurate.

## FUTURE TRENDS

Extracting a set of association rules containing only the most interesting and useful association rules to the end-user is still an important research field. Several solutions addressing this problem have been proposed. These solutions can be classified in two main categories:

- Objective methods that are based on objective interestingness measures, association rule structural properties or hierarchical structures describing relations between items to filter, merge or generalize association rules in order to suppress redundant patterns.
- Subjective methods that integrate the user's beliefs and background knowledge in the domain to suppress uninteresting association rules and select the mot interesting association rules from the user's viewpoint.

An important research topic in this domain is: How can we integrate the different solutions from these two categories in order to select association rules that are the most interesting from both the statistical, the structural and the user's knowledge viewpoints?

Such a method, integrating both subjective and objective pruning criteria, was recently proposed by Chen *et al.* (2008). This method, based on semantic networks to represent user's knowledge, classifies association rules into five categories: Trivial, known and correct, unknown and correct, known and incorrect, unknown and incorrect. Preliminary experiments on a biomedical dataset showed that the reduction can be very important as more than 97 % strong association rules were identified as trivial or incorrect.

This preliminary work show that such methods could greatly improve the usefulness of association rule extraction. An interesting topic in this direction is the development of efficient methods to integrate user's knowledge in the discovery of condensed representations to select the most relevant association rules from the user's viewpoint. This integration could also be done during the visualization phase, by integrating quality measures and user defined constraints for instance.

## CONCLUSION

The frequent closed itemsets theoretical framework was introduced in association rule mining to address the efficiency problem of mining large datasets containing dense or highly correlated data. Several posterior studies showed that this framework is also well-suited to address the problem of redundant association rules that is crucial in most applications. These association rules often constitute the majority of extracted association rules when mining large datasets. They are informatively useless and their suppression is highly desirable to improve the relevance of extracted association rules and simplify their interpretation by the end-user. This chapter focuses on condensed representations and bases for association rules. They are characterized in the frequent closed itemsets framework to show their properties from the theoretical, intelligibility, soundness and informativeness viewpoints.

Condensed representations are reduced sets of association rules summarizing a set of strong association rules. Several condensed representations such as the DG and Proper bases (Pasquier *et al.*, 1999b), the SSMG (Dong *et al.*; 2005), the Reliable Exact Basis (Xu & Li, 2007) and the δ-TARs set (Cheng *et al.*, 2008) were proposed in the literature. These condensed representations are not information lossless since some strong association rules cannot be deduced, with their supports and precision measures, from the rules in the basis. Even if they are not information lossless, they represent interesting alternatives for the presentation to the end-user of a very small set of association rules selected among the most relevant association rules according to objective criteria.

Condensed representations that are information lossless, i.e. from which all strong association rules can be deduced, are called generating sets. Generating sets are defined according to an inference system determining how this deduction is done. A generating set that is minimal with respect to the number of association rules is called a basis. Strong association rules that are deducible by inference are called redundant association rules and a basis contains only non-redundant association rules. A basis covering all itemsets and objects covered by the set of strong association rules is called an informative basis. In other words, each strong association rule must be represented by an association rule of the informative basis. This property is important for the intelligibility of the basis as it ensures that all information are present in the basis and that no computation is required during the interpretation phase.

The minimal antecedent and consequent basis (Zaki, 2000), the Generic Basis (Bastide *et al.*, 2000a) and the IGB Basis (Gasmi *et al.*, 2005) are informative bases defined according to different criteria that correspond to different inference systems. Several studies have shown that these bases are lossless, sound and contain a small number of association rules, that is the minimal number of rules with respect to their inference system. The minimal antecedent and consequent basis was defined outside the scope of the frequent closed itemsets framework. It was defined according to the property that rules with the smallest antecedents are the easiest to interpret for the end-user. By construction, they also have a minimal consequent. The Generic Basis, defined according to the frequent closed itemsets framework, contains rules with the smallest antecedent and the maximal consequent, in order to ease their interpretation and maximize the information in each rule (Mc Garry, 2005). The Generic Basis brings to the end-user all knowledge contained in the strong association rules in a minimal number of association rules without information loss. This basis maximizes the information contained in each association rule and several rules of the minimal antecedent and consequent basis can be summarized by one rule of the Generic Basis. The IGB Basis is an extension of the Generic Basis. It introduces a new kind of rules with an empty antecedent that represent unconditional

co-occurrences of items in the dataset. Then, the association rules of the Generic Basis that can be deduced using these item co-occurrence rules are suppressed to form the IGB Basis. These three informative bases define a compact set of relevant association rules that is easier to interpret for the end-user. Since their size is reduced and as they are generating sets, they also constitute efficient solutions for the long-term storage on secondary memories and the computer-aided management of a set of strong association rules.

## REFERENCES


Agrawal, R., Imielinski, T., & Swami, A.N. (1993). Mining association rules between sets of items in large databases. *Proceedings of the SIGMOD Conference* (pp. 207-216).

Angiulli, F., Ianni, G., & Palopoli, L. (2001). On the complexity of mining association rules. *Proceedings of the SEBD conference* (pp. 177-184).

Armstrong, W. W. (1974). Dependency structures of data base relationships. *Proceedings of the IFIP congress* (pp. 580-583).

Baralis, E., & Psaila, G. (1997). Designing templates for mining association rules. *Journal of Intelligent Information Systems*, *9*, 7-32.

Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., & Stumme, G. (2000a). Mining minimal non-redundant association rules using frequent closed itemsets. *Proceedings of the CL conference* (pp. 972-986).

Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., & Lakhal, L. (2000b). Mining frequent patterns with counting inference. *SIGKDD Explorations*, *2*(2), 66-75.

Bayardo, R. J., & Agrawal, R. (1999). Mining the most interesting rules. *Proceedings of the KDD conference* (pp. 145-154).

Bayardo, R. J., Agrawal, R., & Gunopulos, D. (2000). Constraint-based rule mining in large, dense databases. Knowledge Discovery and Data Mining, *4*(2), 217-240.

Ben Yahia, S., Hamrouni, T., & Nguifo, E.M. (2006). Frequent closed itemset based algorithms: A thorough structural and analytical survey. *SIGKDD Explorations*, *8*(1), 93-104.

Bonchi, F., Giannotti, F., Mazzanti, A., & Pedreschi, D. (2005). Efficient breadth-first mining of frequent pattern with monotone constraints. *Knowledge and Information Systems*, *8*(2), 131-153.

Bonchi, F., & Lucchese, C. (2004). On closed constrained frequent pattern mining. *Proceeding of the IEEE ICDM conference* (pp. 35-42).

Bonchi, F., & Lucchese, C. (2006). On condensed representations of constrained frequent patterns. *Knowledge and Information Systems, 9*(2), 180-201.

Boulicaut, J. F., Bykowski, A., & Rigotti, C. (2003). Free-sets: A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, *7*, 5-22.

Boulicaut, J. F., Jeudy, B. (2002). Optimization of association rule mining queries. *Intelligent Data Analysis Journal*, *6*, 341–357.

Brin, S., Motwani, R., Ullman, J. D., & Tsur., S. (1997). Dynamic itemset counting and implication rules for market basket data. *Proceedings ACM SIGMOD conference* (pp. 255-264).

Brijs, T., Vanhoof, K., & Wets, G. (2003). Defining interestingness for association rules. *International Journal of Information Theories and Applications*, *10*(4), 370–376.

Ceglar, A., & Roddick, J.F. (2006). Association mining. *ACM Computing Surveys*, *38*(2).

Cheng, J., Ke, Y., & Ng, W. (2008). Effective elimination of redundant association rules. *Data Mining and Knowledge Discovery*, *16*(2), 221-249.

Chen, P., Verma, R., Meininger, J. C., & Chan, W. (2008). Semantic analysis of association rules. *Proceedings of the FLAIRS Conference*, Miami, Florida.

Cheung, Y.-L., & Fu, A. (2004). Mining frequent itemsets without support threshold: With and without item constraints. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1052-1069.

Cristofor, L., & Simovici, D.A. (2002). Generating an informative cover for association rules. *Proceedings of the ICDM conference*.

Deogun, J.S., & Jiang, L. (2005). SARM - succinct association rule mining: An approach to enhance association mining. *Proceedings ISMIS conference* (pp. 121-130), *LNCS 3488*.

Dong, G., Jiang, C., Pei, J., Li, J., & Wong, L. (2005). Mining succinct systems of minimal generators of formal concepts. *Proceeding of the DASFAA conference* (pp. 175-187). *LNCS 3453*.

Duquenne, V., & Guigues, J.-L. (1986). Famille minimale d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, *24*(95), 5-18.

Freitas, A. (1999). On rule interestingness measures. *Knowledge-Based Systems*, *12*(5), 309-315.

Carvalho, D. R., Freitas, A. A., & Ebecken, N. F. (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. *Proceedings of the PKDD conference* (pp. 453-461).

Ganter, B., Stumme, G., & Wille, R. (2005). Formal Concept Analysis: Foundations and applications. *Lecture Notes in Computer Science*, *3626*.

Gasmi, G., Ben Yahia, S., Nguifo, E. M., & Slimani, Y. (2005). IGB: A new informative generic base of association rules. *Proceedings of the PAKDD conference* (pp. 81-90).

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, *38*(3).

Goethals, B., Muhonen, J., & Toivonen, H. (2005). Mining non-derivable association rules. *Proceedings of the SIAM conference*.

Hamrouni, T., Ben Yahia, S., & Mephu Nguifo, E. (2006). Generic association rule bases: Are they so succinct? *Proceedings of the CLA conference* (pp. 198-213).

Hamrouni, T., Ben Yahia, S., Mephu Nguifo, E., & Valtchev, P. (2007). About the Lossless Reduction of the Minimal Generator Family of a Context. *Proceedings of the ICFCA conference* (pp. 130-150).

Hamrouni, T., Ben Yahia, S., & Mephu Nguifo, E. (2008). Succinct system of minimal generators: A thorough study, limitations and new definitions. *Lecture Notes in Computer Science*, *4923*, 80-95.

Han, J., & Fu, Y. (1995). Discovery of multiple-level association rules from large databases. *Proceedings of the VLDB conference* (pp. 420-431).

Hilderman, R., & Hamilton, H. (1999). Heuristic measures of interestingness. *Proceedings of the PKDD conference* (pp. 232-241).

Hilderman, R., & Hamilton, H. (2001). Knowledge Discovery and Measures of Interest. Kluwer Academic Publishers.

Hussain, F., Liu, H., Suzuki, E., & Lu, H. (2000). Exception rule mining with a relative interestingness measure. *Proceedings of the PAKDD conference* (pp. 86–97).

Jaroszewicz, S., & Scheffer, T. (2005). Fast discovery of unexpected patterns in data relative to a bayesian network. *Proceeding of the ACM SIGKDD conference* (pp. 118-127).

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A.I. (1994). Finding interesting rules from large sets of discovered association rules. *Proceedings of the CIKM conference* (pp. 401-407).

Kryszkiewicz, M. (1998). Representative association rules and minimum condition maximum consequence association rules. *Proceedings of the PKDD conference* (pp. 361-369), *LNCS 1510*.

Kryszkiewicz, M. (2002). Concise representations of association rules. *Lecture Notes in Computer Science*, *2447*, 187-203.

Lakshmanan, L., Ng, R., Han, J., & Pang, A. (1999). Optimization of constrained frequent set queries with 2-variable constraints. *Proceeding of the ACM SIGMOD international conference* (pp. 157-168).

Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, *184*(2), 610-626.

Leung, C., Lakshmanan, L., & Ng, R. (2002). Exploiting succinct constraints using FP-Trees. *ACM SIGKDD Explorations*, *4*(1), 40-49.

Li, J., Li, H., Wong, L., Pei, J., & Dong, G. (2006). Minimum description length (MDL) principle: Generators are preferable to closed patterns. *Proceedings of the AAAI conference* (pp. 409-414).

Li, J., Tang, B., & Cercone, N. (2004). Applying association rules for interesting recommendations using rule templates. *Proceedings of the PAKDD conference* (pp. 166-170).

Liu, B., Hsu, W., Mun, L.-F. & Lee, H.-Y. (1999). Finding interesting patterns using user expectations. *Knowledge and Data Engineering*, *11*(6), 817-832.

Liu, B., Hsu, W., & Chen, S. (1997). Using general impressions to analyze discovered classification rules. *Proceedings of the KDD conference* (pp. 31-36).

Liu, B., Hsu, W., & Ma, Y. (1999). Pruning and summarizing the discovered associations. *Proceedings of the ACM SIGKDD conference* (pp. 125-134).

Liu, H., Lu, H., Feng, L., & Hussain, F. (1999). Efficient search of reliable exceptions. *Proceedings of the PAKDD conference* (pp. 194-204).

Liu, G., Li, J., & Limsoon, W. (2007). A new concise representation of frequent itemsets using generators and a positive border. *Knowledge and Information Systems*, *13*.

Lu, N., Wang-Zhe, C.-G. Zhou; J.-Z. Zhou. (2005). Research on association rule mining algorithms with item constraints. *Proceedings of the CW conference*.

Luxenburger, M. (1991). Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, *29*(113), 35-55.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*. Cambridge University Press.

Maier, D. (1983). *The theory of Relational Databases*. Computer Science Press.

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases, *IEEE Transactions on Knowledge and Data Engineering*, *5*(6), 903-913.

Ng, R., Lakshmanan, L., Han, J., & Pang, A (1998). Exploratory mining and pruning optimizations of constrained associations rules. *Proceeding of the ACM SIGMOD conference* (pp. 13–24).

Padmanabhan, B., & Tuzhilin, A. (1998). A belief driven method for discovering unexpected patterns. *Proceedings of the KDD conference* (pp. 94-100).

Padmanabhan, B., & Tuzhilin, A. (2000). Small is beautiful: Discovering the minimal set of unexpected patterns. *Proceedings of the KDD conference* (pp. 54-63).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1998). Pruning closed itemset lattices for association rules. *Proceedings of the BDA conference* (pp. 177-196).

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999a). Efficient mining of association rules using closed itemset lattices. *Information Systems*, *24*(1), 25-46.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999b). Discovering frequent closed itemsets for association rules. *Proceedings of the ICDT conference* (pp. 398-416).

Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., & Lakhal, L. (2005). Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, *24*(1), 29-60.

Pasquier, N. (2005). Mining association rules using frequent closed itemsets. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 752-757). Idea Group Publishing.

Pei, J., Han, J., & Lakshmanan, L. V. (2001). Mining frequent itemsets with convertible constraints. *Proceedings of the ICDE conference* (pp. 433–442).

Pei, .J, & Han, J. (2002). Constrained frequent pattern mining: A pattern-growth view. *ACM SIGKDD Explorations*, *4*(1), 31-39.

Pfaltz, J., & Taylor, C. (2002). Closed set mining of biological data. *Proceedings of the BioKDD conference* (pp. 43-48).

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases* (pp. 229-248), AAAI/MIT Press.

Piatetsky-Shapiro, G., & Matheus, C. (1994). The interestingness of deviations. *Proceedings of the KDD conference* (pp. 25-36).

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transaction on Knowledge and Data Engineering*, *8*(6), 970-974.

Silverstein, C., Brin, S., Motwani, R., & Ullman, J. D. (2000). Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, *4*(2), 163-192.

Srikant, R., & Agrawal, A. (1995). Mining Generalized Association Rules. *Proceedings of the VLDB conference* (pp. 407-419).

Srikant, R., Vu, Q., & Agrawal, R. (1997). Mining association rules with item constraints. *Proceedings of the KDD conference* (pp. 67-73).

Stumme, G., Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (2001). Intelligent structuring and reducing of association rules with formal concept analysis. *Proceeding of the KI conference* (pp. 335-350). *LNCS 2174*.

Tan, P., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the ACM SIGKDD conference* (pp. 32-41).

Toivonen H., Klemettinen, M., Ronkainen, P., Hätönen, K., & Mannila, H. (1995). Pruning and grouping discovered association rules. *Proceedings of the ECML workshop* (pp. 47-52).

Valtchev, P., Missaoui, R., & Godin, R. (2004). Formal concept analysis for knowledge discovery and data mining: The new challenges. *Lecture Notes in Computer Science*, *2961*, 352-371.

Wang, K., Jiang, Y., & Lakshmanan, L. (2003). Mining unexpected rules by pushing user dynamics. *Proceeding of the ACM SIGKDD conference* (pp. 246–255).

Xu, Y., & Li, Y. (2007). Generating concise association rules. *Proceedings of the ACM CIKM conference* (pp. 781-790).

Zaki, M.J. (2000). Generating non-redundant association rules. *Proceedings of the KDD conference*.

Zaki, M.J. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, *9*(3), 223-248.

## ENDNOTES

1 Data lines correspond to data rows when the dataset is represented as a data matrix and to transactions when the dataset is represented as a transactional database. Each data line contains a set of items and the set of data lines constitute the dataset.

2 The generic term *object* refers to a data line represented either as a data row or as a transaction.

3 Maximal and minimal itemsets are defined according to the inclusion relation.

4 The set of objects *covered* by an association rule is defined as the set of objects containing the items in the antecedent and the consequent of the rule.

5 A set of inference rules define a procedure which combines association rules in the basis to deduce, or infer, other strong association rules.

6 In the rest of the chapter, statistical measures of precision computed from supports of the antecedent and consequent itemsets of association rule only are noted "precision measures" for simplicity.

# Section VI
# Maintenance of Association Rules and New Forms of Association Rules